Methods in Empirical Prosody Research

Edited by
Stefan Sudhoff, Denisa Lenertová, Roland Meyer,
Sandra Pappert, Petra Augurzky, Ina Mleinek,
Nicole Richter, Johannes Schließer

Offprint

Walter de Gruyter · Berlin · New York

Katrin Schneider, Britta Lintfert, Grzegorz Dogil & Bernd Möbius (Stuttgart)*

Phonetic Grounding of Prosodic Categories

1 Introduction

The phonetic grounding of features, phonemes, and rules or constraints (and, to a lesser extent, rule or constraint interactions) has received increased attention in recent years, and the discussion about phonetic grounding has included aspects such as perception, memory, and motor control in addition to the classical areas of articulation and acoustics; as Pierrehumbert (2000, p. 8) states, "[...] there is no substance-free part of phonology". In our paper we address the phonetic grounding of prosodic categories from the perspective of speech production, perception, and acoustics.

Linguistic categories emerge when the expressions which code them form natural classes. Natural classes are determined by rules. Whenever linguistic expressions are treated as the same by a rule, they form a natural class, and they automatically receive a categorical status.

This classical account of phonological categorization has been extended to prosodic categories as well. Early seminal work in autosegmental phonology has discovered several tune-to-text association rules of a clear categorical status. In his seminal study of Swedish accents, Bruce (1977) presented a set of phonotactic constraints pertaining to pitch accent expressions and focus expressions. These constraints clearly established a categorical status of prosodic categories coding accent and focus in Swedish. However, this categorical status is strongly due to the text side of the tune-to-text association. Although the prosodic expression of the tunes in Swedish has been proven to be highly categorical, its source lies in the linguistic structure it is associated with – the information structure, or the functional sentence perspective, as Bruce called it, following the classical Prague School terminology. The properties of tunes are thus predetermined by the properties of text.

Similarly, Clements and Ford (1979) showed the categorical status of tonal categories in Kikuyu, which not only adhere to synchronic processes but are

[*] We wish to thank two anonymous reviewers for their helpful comments and suggestions on a previous version of this paper, and Vincent van Heuven for the fruitful discussion of the experimental methods and results in the context of the perception of boundary tones. All errors remain ours.

also supported by sound laws. In Kikuyu, the categorical structures of tonal expressions very much depend upon the linguistic association site. Each morpheme contributes a tone to the tone melody (the tune) of a word, but general association rules generate tunes where morphemes are not necessarily associated with their underlying tones (Goldsmith, 1990, pp. 11–14).

The strong evidence for categorical systems in prosody always involves the properties of tune-text association like the cases illustrated in Kikuyu. The categorical status of tunes alone is much more controversial. Actually, there is a strong research tradition claiming that prosodic categories, and particularly the tunes, are quite different from all other linguistic categories. Whereas linguistic categories are compositional, and their meanings can be discerned only by logical analytical reasoning and decomposition, the prosodic categories are holistic, gestalt-like, and have immediate iconic reference. Helmholtz (1863) coined the term "unbewusstes Schließen" ('unconscious reasoning') when referring to the interpretation of the prosodic cues. However, there are now reasons to assume that the iconic, non-symbolic aspects of prosody interpretation are strongly overemphasized (Dogil, 2003). The experimental methodology that we present in this paper aims to provide evidence that elements of the tonal structure are perceived categorically.

The more general question, however, is how these categories emerge even in languages where there are no hard rules forcing prosodic expressions into natural classes. We follow a theoretical perspective according to which prosodic categories emerge as invariant regions in the perceptual space of speakers/hearers of a language (Dogil and Möbius, 2001; Möbius and Dogil, 2002). This process is depicted in Figure 1.

We argue that prosodic cues like pitch, loudness, duration and voice quality are initially processed at prosodic landmarks in the speech signal. These are vowels in tone languages, accented units in pitch-accent languages, and stressed syllables and final syllables (boundaries) in intonational languages. The values extracted at the landmarks establish the first approximation of a pertinent prosodic category. The representation is highly underspecified; the incremental process of specification proceeds by considering the context of the estimated category. The phonetic context further specifies the alignment with phonetic categories (e.g., early and late alignment points) and micro-prosodic features (influence of stops, laryngeals, etc.).

The linguistic context adds the morpho-syntactic specification to the category and takes cliticization and syntactic structure into consideration (Cinque, 1993). The discourse context places prosodic features in a semantic perspective and brings givenness, topic, and focus into play (Dogil et al., 1997; Büring, 1997). Finally, the characteristics of the speaker's voice, e.g., voice quality parameters or duration z-scores (see Section 3.3.2) also contribute to the more specific representation of a presumed prosodic category.

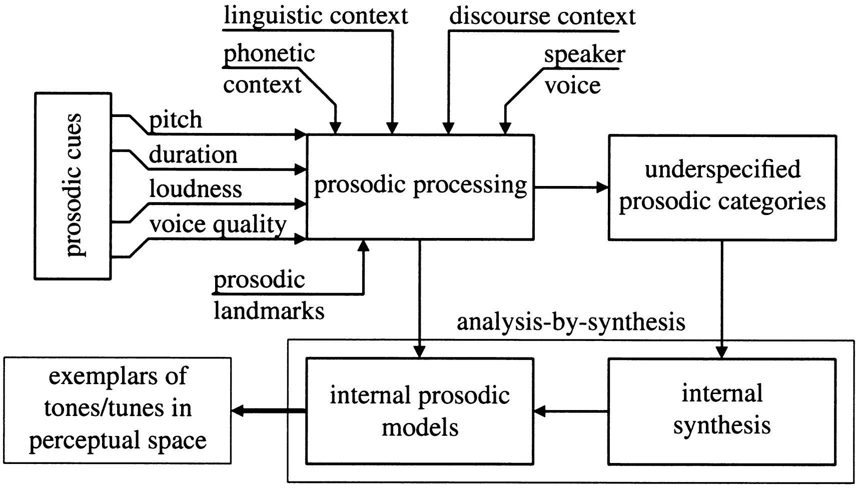

Figure 1: Prosodic categories emerge as the result of an internal analysis-by-synthesis process and are defined by category-specific exemplar clouds in the perceptual space.

The upper part of Figure 1 constitutes a more refined view of the autosegmental tune-to-text association principle. However, it is only a part of the prosodic recognition process. The emergence of a prosodic category is reached by the application of an internal analysis-by-synthesis process illustrated in the lower part of Figure 1. This process takes the underspecified category and sends it through a loop in which the category is internally synthesized by the hearer (the right box of the analysis-by-synthesis loop). Then the result of the synthesis is internally compared with the original signal in the original context. Such a comparison can be performed only in the phonetic space that is neutral to the hearer's and the speaker's articulatory-acoustic space. We follow Guenther, Perkell and others (Guenther et al., 1998; Perkell et al., 2000; Dogil and Möbius, 2001) and define the internal prosodic models in a perceptual reference space.

The analysis-by-synthesis loop is repeated as often as the category is encountered. This means that frequency of experience is decisive for the establishment of a prosodic category. The cues which have a high frequency of occurrence in a specific context will be represented by a large number of internal exemplars in the hearer's perceptual reference space. These exemplars develop categorical properties and exert magnet effects. We will therefore argue that phonetic categories emerge from probability distributions corresponding to regions in the parametric phonetic space.

One aspect of phonetic grounding pertains to the question of whether bottom-up factors exist that favor the emergence of certain categorizations over others. Empirically, the utilization of the phonetic space is not uniform; this effect has been studied in great detail with respect to the vowel space. Stevens (1989) has argued that the relationships between articulatory parameters and their acoustic and auditory responses are nonlinear and that both articulatory-acoustic and acoustic-auditory relationships are quantal.

An important question is whether such nonlinearities exist in the prosodic domain as well. Section 2 therefore presents methods for testing whether there are also bottom-up factors that favor the emergence of prosodic categories. We suggest an array of perceptual tests that comprises the classical paradigm of Categorical Perception (Liberman et al., 1957; Repp, 1984) and the concept of the Perceptual Magnet Effect (Kuhl, 1991). Both paradigms can be applied to segmental as well as prosodic categories and it is assumed that both produce similar categorical results.

The second aspect of phonetic grounding concerns the mapping of prosodic categories onto continuous acoustic parameters of the speech signal. In Section 3, a computational approach to modeling the phonetic cue implementation of prosodic categories is proposed which serves to describe the structure of prosodic categories by revealing their parametric acoustic dimensions. We will demonstrate that probabilities based on frequencies of occurrence are essential for this computational model and, following Pierrehumbert (2003), we expect probability distributions also to be a key ingredient of a model of phonetic grounding in the prosodic domain.

2 Experimental evidence for prosodic categories

In this section we discuss methods for testing the categorical status of elements of the prosodic structure: the classical paradigm of Categorical Perception (Repp, 1984), and the concept of the Perceptual Magnet Effect (Kuhl, 1991).[1] Our own experiments on the categorical perception of boundary tones in German show that discrimination is generally better than predicted by the identification function. This finding seems to indicate that the observed perceptual patterns are more in line with the perceptual magnet effect than with the classical categorical perception. We focus on the experimental design, the stimulus generation, and how to evaluate and interpret the experimental results; our own experiments are presented to illustrate these methodological issues in detail.

2.1 Testing Categorical Perception

An experiment testing for Categorical Perception (CP) involves two perceptual tasks: an identification test where stimuli have to be assigned to given categories, and a discrimination test where pairs of stimuli are evaluated as to whether they

[1] Other experimental designs that may be suitable to test for the existence of categories, such as the imitation task, which involves the perception and the production of stimuli, are not discussed in this paper.

consist of identical or different stimuli. Perception is considered to be categorical if the category boundary (the crossover between the categories) found in identification corresponds to the peak in discrimination, i.e. to the point where most stimuli are distinguished.

The classical paradigm of categorical perception was developed on the basis of the perception of plosives (Repp, 1984). It is not clear if the definition of CP as it stands can be adopted to intonation. Acoustically, intonational units behave somewhat similarly to vowel phonemes in that they are encoded over relatively long time intervals. Thus, in the case of CP, intonational units may be expected to show a plateau in discrimination rather than a peak, just as vowels do. The categorical perception paradigm has only rarely been applied to the prosodic domain, with a few notable exceptions.

House (1996) has shown that in areas of spectral change, pitch movement is perceived as a level tone. The perception of a pitch movement depends on its temporal alignment and spectral information. Only if the pitch movement is located in an optimal location, i.e. in a region of little spectral change, can it be perceived categorically. Kohler (1987, 1990) demonstrated categorical changes in the perception of early vs. medial F_0 peaks. Ladd and Morton (1997) carried out a series of experiments to test the idea that there is a categorical difference between normal and emphatic accent peaks in English, rather than a continuum of gradually increasing emphasis. Based on their results they suggested that the normal vs. emphatic distinction may be categorically interpreted even though it is not categorically perceived. On the other hand, Remijsen and van Heuven (1999) reported categorical perception of Dutch boundary tones.

In our experiment we adopted Remijsen and van Heuven's experimental design and focused on a small part of the German intonation system: the perception of boundary tones. Phonologically, the categorical status of boundary tones is widely undisputed, although, with the exception of the experiment by Remijsen and van Heuven (1999) for Dutch, it has never been tested for any other language. We therefore tested the categorical status of low (L%) and high (H%) boundary tones in German, typically representing statements and certain types of questions, respectively.

Intonational phonology posits that the boundary tones L% and H% are members of a primary phonological opposition, which entails that they distinguish between different meanings, including semantic and pragmatic interpretations represented by different sentence modes, such as "statement vs. question".[2]

[2] One reviewer pointed out that a design that uses out-of-the-blue sentence stimuli provides no control of the semantic or pragmatic interpretations by the listener. We agree, but we argue that providing an explicit context to the stimulus will constrain too strongly and bias the possible interpretations.

2.1.1 CP test preparation

To test for categorical perception a stimulus continuum has to be created that covers the phonetic space comprising the target categories. In the case of the categorical status of the low and high boundary tones in German, the text of the test stimulus had to be a sentence that was syntactically ambiguous between statement and question, the contrast being encoded by the fundamental frequency (F_0) within the last syllable of the utterance. As the F_0 contour is the primary correlate of intonation, intensity and duration were kept constant in order to isolate the perceptual effect of F_0. Our intention was to test the perception not in synthesized but in natural speech. Synthesized stimuli are often judged by listeners as sounding unnatural.

To obtain a syntactically ambiguous sentence in natural speech, a professional male speaker of German read several dialogs with normal intonation. These utterances were recorded in the anechoic chamber at the Institute of Natural Language Processing (IMS) at the University of Stuttgart. From these dialogs the test sentence "Steht alles im Kochbuch" ('It's / Is it all in the cookbook') was selected, which satisfied all conditions required for manipulation:

- Presented in isolation, i.e. without any surrounding context, it is syntactically ambiguous between the sentence modes statement and question. In this situation, subjects can only rely on intonational information to distinguish between statement and question, and they do, as our results show. In German, subjects perceive a low boundary tone as a statement and a high boundary tone as some type of question.[3]
- The sentence-final syllable is unaccented. This property avoids the presence of two linguistic functions (sentence mode and accenting) in the same syllable, which would make it difficult to modify just one of these functions (sentence mode).
- The sentence-final syllable must not contain a schwa vowel, because this sound is often very short or sometimes not even realized, so that the manipulation may lead to no audible change or produces very unnatural sounding examples.
- The F_0 contour of the test sentence's final syllable forms a plateau-like contour ending on a medium fundamental frequency level in the speaker's pitch range (128 Hz). By using this intermediate pitch level as a starting value for F_0 manipulations, we were able to create an acoustic F_0 stimulus continuum covering the L% and H% ranges by means of the PSOLA algorithm (Moulines and Charpentier, 1990) without introducing audible distortions in the acoustic stimuli.

[3] One reviewer pointed out that the chosen sentence may be syntactically biased in favor of an interrogative sentence mode because, in German, V1 sentences are typically used for Yes/No questions. This is true; in fact, we would like to add that this syntactic bias may or may not have been compensated for by the fact that declaratives are more frequent in actual language usage than interrogatives. Future experiments should attempt to use sentences that avoid such biases.

We only manipulated the last syllable of the test sentence. Intonational theory posits that the boundary tone is associated with the edge of the phrase and aligned with the phrase-final syllable (e.g. Féry, 1993). It is known that changing the F_0 course of the last syllable is sufficient to achieve the perceptual effect of different boundary tones.[4] Manipulation of the test sentence was based on the average rising or falling F_0 slope within the last syllable in questions and statements produced by the speaker.

We analyzed all sentences produced by the speaker and calculated that he produced a typical statement (L%) with an average fall of 50 Hz within the final syllable, and a typical question (H%) with an average rise of 100 Hz within the final syllable. For our test sentence, this resulted in 228 Hz for the question (H%) and 78 Hz for the statement (L%). These values restrict the range of the test continuum.

The continuum consisted of 11 equally-sized steps (including both extremes) to meet two criteria: first, the step size should be neither too large nor too small to pose an adequate challenge to discriminability; and second, to obtain a number of stimuli that allows repeated presentations for the purpose of statistical analysis while keeping the test duration reasonably short.

For manipulation we used the ERB scale (Equivalent Rectangular Bandwidth). This frequency scale is considered to be the most satisfactory psychophysical transformation of pitch intervals in human speech (Hermes and van Gestel, 1991). The ERB scale is linear below 500 Hz and nearly logarithmic above this frequency. ERB values are calculated from Hertz values according to equation 1.

$$ERB = 16.7 * log\left(1 + \frac{F}{165.4}\right) \tag{1}$$

Using 11 manipulation steps resulted in a step size of 0.35 ERB in our experiment (Figure 2). After calculating the ERB values of the boundary tones for each of the 11 manipulation steps, the values were transformed back into Hertz values. The F_0 contours of our test stimuli were resynthesized by means of PSOLA (Moulines and Charpentier, 1990).

2.1.2 The CP experiment

Participants. Participants in a CP experiment are native speakers of the pertinent language without hearing deficits. In our experiment testing for categorical perception, 24 native German listeners, 13 male and 11 female, participated on a voluntary basis. The experiment was Web-based and self-paced.

[4] Manipulation of natural speech in order to obtain different presentation stimuli which are themselves natural is not contradictory in itself. First, we manipulated only one parameter (F_0), and second, within certain limits, the PSOLA pitch modification technique produces stimuli whose acoustic quality is indistinguishable from that of natural ones.

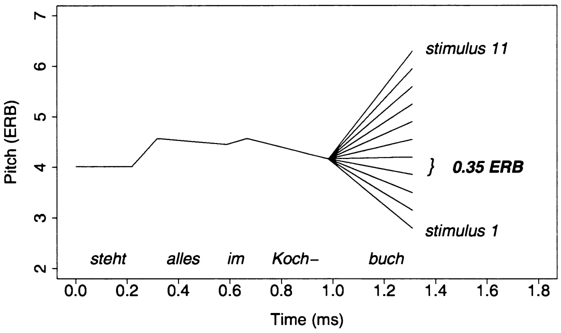

Figure 2: Manipulation of the sentence-final syllable of the test sentence in steps of 0.35 ERB to obtain 11 different stimuli

Experimental procedure. The subjects listened to the stimuli via headphones in order to minimize ambient noise. The volume of the stimuli was set at a comfortable level at the beginning of the experiment. Subjects had to choose one of the response alternatives after listening to a stimulus only once, before the next stimulus was presented.

Identification. During identification, listeners had to classify the stimulus as either a *question* or a *statement*. The stimuli were repeated 10 times and presented in random order. The results pooled for all participants show clear s-shaped curves, which is a typical result for categorical perception (Figure 3). A full crossover, i.e. a rise from less than 20% to more than 80% identification of one category, was reached by each subject within only two steps. There were differences between participants in the exact location of the category boundary. Therefore, the full crossover from *statement* to *question* averaged over all participants did not occur within two but within five steps. This is a first hint that categorical perception in intonation may not be as easy to demonstrate as in the case of segmental phenomena.

Within each category, identification is on a constant, high level. These findings were subsequently verified and confirmed by a proportion test, indicating that CP was indeed found.

Discrimination. In the discrimination tests, subjects listened to stimulus pairs that were either identical (AA sequence) or one step apart in the continuum (Remijsen and van Heuven, 1999). In the latter case, the second stimulus was either higher (AB sequence) or lower (BA sequence) than the first one. The classical definition of categorical perception predicts that discrimination between adjacent stimuli within a category is nearly impossible, and discrimination between stimuli belonging to different categories is very good. This implies that those stimulus pairs which include stimuli that straddle the category crossover

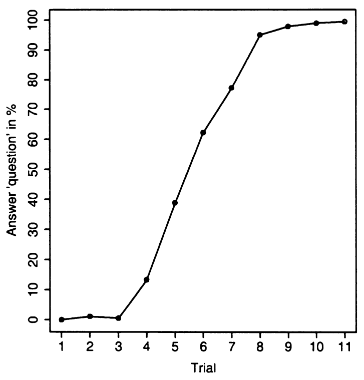

Figure 3: Results of the identification test averaged over all participants

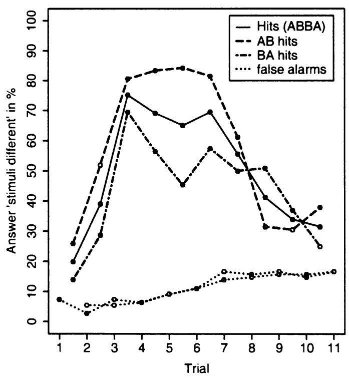

Figure 4: Results of the discrimination test averaged over all participants

found in identification are easy to distinguish. In contrast, stimulus pairs consisting of stimuli that belong to the same category are expected to be hard to distinguish.

The silent interstimulus interval (ISI) within each stimulus pair was set to 500 ms. ISI used in the research literature range from 250 ms to 1.5 s. Iverson and Kuhl (1995) demonstrated that the ISI only has an influence on the reaction times of the subjects but no influence on the perception results. We therefore used an ISI that is not too short to avoid that the second sentence starts very shortly after the first one has ended, which may be confusing in sentence perception. If the ISI is too long, the impression of the first test sentence may fade before the subject can rate whether the stimuli are different or identical. Batliner (1989) reported an ISI of 500 ms as useful for testing sentence intonation perception.

In our discrimination test, all pairs were repeated 6 times. This repetition number was chosen taking into account two facts: first, a certain number of results per stimulus pair is necessary for statistical analysis; and second, more repetitions would result in a higher test duration and therefore a decrease in listeners' attention. In the literature, there is general agreement that a perception test should last no longer than one hour. However, listening to the same kind of stimulus for one hour seemed very tiring to us. We therefore reduced the maximum test duration to 45 minutes and split up discrimination into two subtests, each consisting of 3 repetitions of stimulus pairs. Our subjects had to complete three subtests, viz. one identification test and two discrimination tests, with a break of at least one hour between any two subtests.

The typical curve of the discrimination function indicating categorical perception shows a high peak around the trial where the identification crossover between the two categories takes place. The results obtained in our discrimination tests show two separate peaks connected by a high plateau with a mean correctness of about 70% (Figure 4). A proportion test confirmed that the values forming the plateau belong to the same range of values.

Whenever stimuli are presented in pairs, it has to be taken into account that the order of presentation might have an effect (Schiefer and Batliner, 1991): participants are more successful in discriminating one stimulus order compared to the reverse order. Remijsen and van Heuven (1999) and Schneider and Lintfert (2003) found that subjects were more successful in discriminating stimulus pairs when the second stimulus had the higher final pitch (AB sequence). What is important in the discrimination results is the fact that, independently of the order of presentation, all results are clearly above the curve of false alarms. False alarms are stimulus pairs that are marked as being *different* although they consisted of identical stimuli. If the false alarm rate is very high, one can conclude that the participants failed to carry out the test successfully. The false alarm rate during our experiment was very low, which confirms that the step size during manipulation was not too small, and that the test results can be used for statistical analyses.

2.1.3 Interpreting the results of a CP experiment

Categorical perception implies that the category boundary of the identification function correlates with the peak of the discrimination function. The shape of the discrimination curve can be predicted by applying the so-called Haskins formula (2) to the identification results.

$$P(C) = 0.5 * [1 + (p1 - p2)^2] \tag{2}$$

The comparison of the predicted and the obtained curve forms the basis for deciding whether the contrast can be interpreted as being perceived categorically. In our results, the predicted curve shows a plateau which corresponds to the plateau obtained in the discrimination results. This is yet another piece of evidence for CP in the boundary tones of German. But in our discrimination results the crossover varies between participants (see Section 2.1.2). In this case, according to the theory of CP, the individual peaks in discrimination will vary as well (Remijsen and van Heuven, 1999). Therefore it was necessary to evaluate the individual crossover points. We interpolated a straight line through the values of those trials between which the category switch occurred and calculated the exact crossover for each participant. A regression analysis with these crossover points as the predictor variable and the individual peaks in discrimination (averaged over stimulus orders) was performed, which led to a low but significant correlation for our results.

In summary, we found categorical perception of the boundary tones L% and H% in German. Evidently, participants are significantly better at discriminating stimuli at the perceived category boundary than within categories. However, our results show a plateau in discrimination which correlates significantly with a broad category crossover in identification. This is neither consistent with continuous perception, where all stimulus pairs are discriminated at an equal level of correctness, nor with CP. This finding seems to support the doubts that have been raised as to whether intonational categories can be described by CP in the classical sense (Massaro, 1998). Maybe the definition of CP has to be reformulated for intonation.

Another possible interpretation of the results is that a third category is present between L% and H%. One common linguistic principle is the use of minimal effort for achieving the desired result. Therefore, there should be no wasted space between adjacent categories. If L% and H% are indeed adjacent categories, the identification function should be steeper and the discrimination function should have a narrower plateau than in our results. Taking this into account, we suspect a third category to be hidden in our results (cf. Repp, 1984), a category that was not one of the offered responses.[5] Evidence for such an additional category was already present in the results of the individual listeners. Although there was always a steep identification function, the individual discrimination curves showed lower plateau-like contours or a smaller peak beside the highest one. This hypothesized third category might be *continuation*, which is supposed to lie on an intermediate level between *statement* and *question*. But *continuation* might merely be another term for a feature discussed below: terminality.

Intonation conveys several types of information simultaneously. It is possible that intonational categories cannot be distinguished from one another with respect to one single feature but only to a combination of features. For our test sentence, a low boundary tone (*statement*) is always considered to be turn-terminal, whereas a non-low boundary tone can be either terminal or non-terminal. Subjects in our experiment had to assign all stimuli, including those representing the putative non-low and non-terminal ones, to one of the two offered categories. Thus, the non-low and non-terminal boundary tone might have been responsible for the observed broad crossover and the discrimination plateau. It is important to note that the broad crossover is not the result of the listeners' random assignment of stimuli but the result of between-subject variability in terms of the location of the category switch. Individual subjects showed sharp category crossovers. Apparently, listeners used one of the following two strategies: either everything that was not a clear *statement* (low and terminal) was a *question*, or everything that was not a clear *question* (high and terminal) was a *statement* (cf. also van Heuven and Kirsner, 2004). This hypothesis was taken into account in the design of the identification test of the experiment testing for the perceptual magnet effect described in the next section.

[5] We offered the subjects only two possible answers in this experiment because we tested only for two boundary tone categories. Our results suggest that the experiment should be repeated using three possible answers for the identification test. In fact, in the subsequent PME experiment (see Section 2.2.2), three answer alternatives were offered for precisely this reason.

2.2 Testing the perceptual magnet effect

Kuhl (1991) demonstrated that newborn infants can discriminate all sounds of the world's languages. But this ability decreases as linguistic knowledge of the child's ambient language increases. For this phenomenon, Kuhl introduced the concept of the native language perceptual magnet effect (PME). The concept of PME states that there may be differences in the discrimination ability within each category because each category consists of a perceptual magnet, represented by a prototype (P). A prototype of a category is the realization that best matches all features of the category. Perception is warped around the prototype: a prototype perceptually attracts its immediate neighbors, i.e. the perceived distance between the prototype and its neighbors is reduced. The prototype has the effect of a magnet for its neighbors, which results in reduced discrimination sensitivity. The immediate neighbors of the prototype are hard to distinguish from the prototype itself. This is not the case between a nonprototype (NP) and its immediate neighbors. These observations contrast with the predictions made by the paradigm of categorical perception.

The perceptual magnet effect is a mechanism which influences phonetic perception by language experience. Once a speaker has fully acquired his own language, he has developed exemplar clouds in memory for all categories of this language. Each category consists of many perceived realizations that have been collected up to this point. Perception decides which and how many categories will be developed and where they are located in the perceptual space. Good instances of a category are easier to determine and easier to remember than other members of the category. They leave better traces in memory than poorer instances, and good instances of a category will be used more often during the production of instances of the category.

Categories may or may not have perceptual magnets. Testing for the presence of a perceptual magnet involves three subtests: first, an identification task; second, a goodness rating; and third, a discrimination task. These three subtests are described below. They have to be carried out in this particular order because the results of the preceding test render the input stimuli for the subsequent test.

2.2.1 PME test preparation

Several steps of the PME test are identical to those in the CP test. In our PME experiment we used the same test sentence as in the CP experiment since the selection criteria were the same. The 11-step stimulus continuum with a step size of 0.35 ERB, as created for the CP experiment, was expanded to a 20-step continuum by producing stimuli with the same step size below and above the typical L% and H% contours, respectively. Further stimuli were produced as long as they sounded natural, which was evaluated by several listeners. In the new stimulus continuum the lowest boundary tone has an F_0 value of 35.3 Hz (= 1.4 ERB) and the highest boundary tone has an F_0 value of 337.5 Hz (=

8.05 ERB). With this extension of the stimulus set we ensured that the presumed perceptual magnets of each boundary tone category were included. The F_0 contours were resynthesized by means of PSOLA (Moulines and Charpentier, 1990) and the stimuli were numbered from 1 (lowest boundary tone) to 20 (highest boundary tone) (Figure 5).

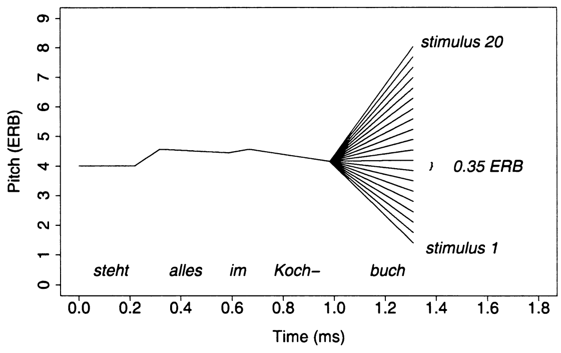

Figure 5: Manipulation of the sentence-final syllable of the test sentence in steps of 0.35 ERB to obtain 20 different stimuli for the PME experiment

2.2.2 The PME experiment

Participants and experimental procedure. The criteria for the selection of subjects and the experimental procedures were the same as in the CP test. A total of 21 subjects (10 male and 11 female) participated in the PME test.

Identification. The identification test procedure was the same as in the identification test for categorical perception. Subjects were asked to classify each stimulus as a member of one of the given categories. In contrast to the CP experiment, three categories were given as possible answers: *question*, *statement*, and *neither question nor statement*. The third category was to be used when the subject was not sure to which of the two main categories the stimulus belonged. The intention was to ensure that the two boundary tone categories found in the CP experiment included only those stimuli that were indeed perceived as members of the respective category by the majority of the participants.

Results varied slightly between individual subjects. Within the two boundary tone categories the identification rates were above 90%, whereas the category *neither question nor statement* had an identification rate of at most 55%. This seems to be counterevidence for the existence of a third category between the low and the high boundary tone in German (cf. Section 2.1).

In our experiment, the category *statement* has more constant identification rates than the category *question*. This results from the identification from 4 subjects who classified the three highest stimuli as belonging to the category *neither question nor statement* because these stimuli sounded unnaturally high to them.

Goodness rating. Based on the results of the identification, separate goodness rating tasks have to be carried out for the two categories *statement* and *question*. To ensure that only those stimuli are included in each of the two categories that are indeed members of the pertinent category, a stimulus was accepted for the goodness rating if it was identified as a member of the pertinent category by more than 75% of the subjects. This requirement was met by stimuli 1-7 and 14-20. Two proportion tests verified that the stimuli in each of the two sets did not differ significantly in their identification rates; they were therefore included in the goodness rating task for the categories *statement* and *question*, respectively.

Listeners were asked to label the quality of each stimulus on a pre-defined scale. We used a scale from 1 (*very bad exemplar of this category*) to 7 (*very good exemplar of this category*). Prior to the rating for each category there was a training session to acquaint the listeners with the range of the stimuli. During training, all different stimuli that had to be evaluated in the real test situation were presented, and the participants could listen to these stimuli as often as they wanted to. During the test for each category, the stimuli were repeated 10 times and presented in random order. Listeners had to choose one of the given labels (1 to 7) before they could listen to the next stimulus. Label corrections were not possible.

For the *statement* category, we found only slight individual differences in the rating of the stimuli. There are two ways of obtaining the prototype and the nonprototype for the *statement* category. The first is to calculate the total rating that each stimulus received from all subjects. The stimulus with the highest rating is the prototype and the stimulus with the lowest rating corresponds to the nonprototype. But this procedure does not take into account that the participants differed in their use of the rating scale. The second possibility is to calculate the stimulus with the highest and the one with the lowest rating for each subject, which results in two sets, one set of individual prototypes and one set of individual nonprototypes. The median of the first set corresponds to the prototype of the *statement* category (P_S), and the median of the second set corresponds to the nonprototype of the *statement* category (NP_S). We used the latter procedure.

For the category *question* there were greater individual differences in the ratings of the stimuli than in the *statement* category. It was more difficult for all subjects to decide between good and not-so-good exemplars of this category than it was for the *statement* category. The method of calculating the *question* prototype (P_Q) and the *question* nonprototype (NP_Q) was the same as for the *statement* category. The difficulties in obtaining a clear prototype in the *question* category may be viewed as a first cue that there is no perceptual magnet in this category.

Discrimination. In the discrimination test pairs of stimuli have to be evaluated as consisting of *identical* or *different* stimuli. In contrast with the discrimination test in the CP experiment, the stimuli within a pair in the PME experiment need not be immediate neighbors in the stimulus continuum. One stimulus of each sentence pair is always either the prototype (P) or the nonprototype (NP) of the category, the other stimulus is either an immediate neighbor of P or of NP, or a neighbor that is two, three or even more steps away. If the discrimination ability around P is reduced and the discrimination ability around NP is not, the perceptual magnet effect is assumed to be present in the examined category.

Discrimination was tested separately for each category. We used two different test designs: first, a random test design, in which the pairs including the prototype and the pairs including the nonprototype were randomly mixed; second, a block design, in which the pairs including the prototype were tested separately from those including the nonprototype. Furthermore, in the second test design, the pairs including P could be presented before the pairs including NP (P-first block design) and vice versa (NP-first block design) to test for the effect of the order of presentation. Each subject participated in only one test design and the assignment of subjects was decided randomly.

Because the number of stimuli was too large for one test session for each category, discrimination was split up into two tests for the *statement* and two tests for the *question* category. All tests had to be performed with a break of at least one hour between any two subtests.

Again, a training session before each discrimination test allowed the subjects to become acquainted with the stimuli and the differences within the pairs. All possible stimulus pairs were presented during training. The subjects had to listen to the stimulus pairs and were asked to decide if the stimuli in the pair had been *identical* or *different*. They were told whether their answer was correct after each training trial. During the tests no feedback was given.

Signal Detection Theory. Although the discrimination task involved different test designs, no effect from the order of presentation was found in the results. However, results of the individual subjects differed not only in their hit rates, i.e. how many pairs they correctly recognized as including different stimuli, but also in their false alarm rates, i.e. how many pairs they wrongly recognized as including different stimuli. This is in line with Signal Detection Theory (SDT) (Wickens, 2002; Heeger, 2003), which takes into account that the attitude of the listeners toward the test influences their results. According to SDT, listeners who share the same perceptual pre-condition (identical auditory threshold) can produce different results in a perception test because they use *response criteria* of different sizes. In any trial, the answer of the observer is YES if the evidence for the signal is larger than a value known as the *response criterion* λ_{Center}, and NO when it is smaller than this value, which implies that the number of hits and false alarms depends on this criterion. λ_{Center} is calculated by taking into account the z-transformations (Gaussian distribution) of the hit rate (h) values and the false alarm rate (f) values, as the following equations (Wickens, 2002) show:

$$h = \frac{\text{Number of hits}}{\text{Number of signal trials}} \tag{3}$$

$$f = \frac{\text{Number of false alarms}}{\text{Number of noise trials}} \tag{4}$$

$$\lambda_{Center} = -0.5 * (Z(f) + Z(h)) \tag{5}$$

Discrimination of the statement category. In our experiment we found for the category *statement* that the hit rates of the stimuli in the immediate vicinity of the prototype $\mathrm{P_S}$ differed significantly in their mean values from those in the immediate vicinity of the nonprototype $\mathrm{NP_S}$: there were significantly more hits in the $\mathrm{NP_S}$ environment. With respect to λ_{Center}, there was also a significant difference between the $\mathrm{P_S}$ and the $\mathrm{NP_S}$ vicinities.

The values for λ_{Center} are significantly higher around $\mathrm{P_S}$ than around $\mathrm{NP_S}$ (Figure 6). Post-hoc tests confirmed that the immediate neighbors of $\mathrm{P_S}$ differ significantly in their values for λ_{Center} from all other stimuli either in the surrounding of $\mathrm{P_S}$ or that of $\mathrm{NP_S}$. Another correlation was found for the response criterion and the results of the goodness rating task for the $\mathrm{P_S}$ environment, with the response criterion increasing with better goodness rating values. We found reduced discriminability in the vicinity of $\mathrm{P_S}$ but not in the vicinity of $\mathrm{NP_S}$. This is exactly what PME assumes: perception is warped around P but not around NP. Thus, we found strong evidence for PME for the *statement* category.

Discrimination in the question category. There were no significant differences between the hit rates of the immediate environment of $\mathrm{P_Q}$ and that of $\mathrm{NP_Q}$ for the *question* category (Figure 7). The same holds for the response criterion λ_{Center}: both surroundings do not differ significantly in their λ_{Center} values. But as in the category *statement*, there is a correlation between λ_{Center} and the trials. The result indicates that the discrimination ability around the supposed prototype of this category is worse than around the nonprototype, but this difference is found only with the second neighbor of $\mathrm{P_Q}$ and $\mathrm{NP_Q}$, respectively, and it is far from reaching significance. We conclude that there is no evidence for a magnet effect in the *question* category in German.

2.2.3 Discussion of the results of the PME tests

The results of our experiments show that a perceptual magnet exists for the *statement* category in German. There were only slight differences between the subjects in the results of the identification, the goodness rating and the discrimination for this category. We therefore conclude that the *statement* category with its

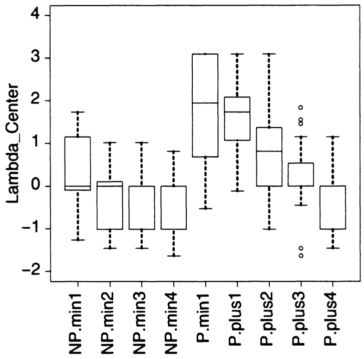

Figure 6: λ_{Center} for the statement category differs significantly between NP and P trials.

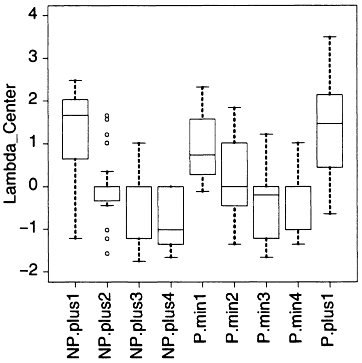

Figure 7: λ_{Center} for the question category shows minor differences between NP and P trials.

low boundary tone is well established in German and has clear limits and contours. We successfully determined a prototype of this category, which shows a very low F_0. As this prototype was almost the lowest one in the stimulus continuum and was still accepted as sounding natural, we suppose that the most important cue for the perception of *statement* is its low boundary tone.

For the *question* category there are no such clear results. There was no evidence for a perceptual magnet in this category. This finding may be explained in different ways. First, there might be a third boundary tone, possibly *continuation* (%). This boundary tone would differ from L% in the height of F_0, i.e. it is non-low, and maybe from H% in the height of F_0 as well, but definitely in being non-terminal. A non-low and non-terminal boundary tone signals that the speaker will continue his turn, whereas a high and terminal boundary tone (H%) signals that the listener may take the turn. Therefore it might be the case that the better discrimination ability in the *question* category is a consequence of discriminating within a perceptual space in which two boundary tones are present, possibly comprising two perceptual magnets. This assumption might explain why the discrimination values for the immediate NP_Q neighbor are as bad as those for the immediate neighbors of P_Q: P_Q corresponds to the *question* prototype (as we supposed) but NP_Q corresponds to the continuation prototype (and not to the nonprototype of the *question* category as we supposed). However, we did not find clear evidence for the existence of a third boundary tone category in German in the identification task.

The second explanation uses arguments from Exemplar Theory (see Section 3). If each category emerges from all the exemplars that the listener perceives, then the definition of a prototype of a category states that this is the location in the exemplar cloud with the highest exemplar density (Lacerda, 1995).

Discrimination is almost impossible there because the exemplars are so close and therefore similar to each other. Precisely at this point, goodness ratings are expected to be maximal. In fact, this is the case for the category *statement*. The goodness ratings for the *question* category showed that several subjects did not accept the highest stimuli as good exemplars of this category. These subjects may have developed different exemplar clouds for the category *question*, possibly due to regional differences.

The emergence of category-specific exemplar distributions and an approach to a computational mapping between the categories and their phonetic substance is addressed in the following section.

3 Emerging prosodic categories

In this section we sketch a model which rests on the hypothesis that categories emerge from probability distributions corresponding to regions in the parametric phonetic space. Key ingredients of this model are several of the assumptions put forward by Guenther and Perkell and their colleagues (Guenther, 1995; Guenther et al., 1998; Perkell et al., 2000; Guenther, 2003) as well as predictions made in the context of Exemplar Theory (Johnson, 1997; Pierrehumbert, 2001a). We also present an experiment designed to test these hypotheses and to provide evidence for the model's validity.

3.1 Exemplar-theoretic model of categorization

The internal analysis-by-synthesis procedure outlined in Section 1 leads to the establishment of exemplars of linguistic units, which are stored in long-term memory together with all the context information available and a category label (Lacerda, 1995; Johnson, 1997; Pierrehumbert, 2001a). The accumulation of exemplars of a given category implicitly defines a multidimensional region in the perceptual reference space. The size and structure of each region are functions of the frequency of usage (viz. synthesis for the purpose of verification) of exemplars and their variability, which in turn depends on contextual variations. There is a unique phonetic target region in the perceptual reference space for each (segmental and prosodic) category of speech in a given language (Perkell et al., 2001; Guenther, 2003).

Frequencies of occurrence, or frequencies of experience, play a central role in the cognitive system of language (Pierrehumbert, 2001a). Probability-related knowledge of the phonetic space must be built up during language and speech acquisition according to the phonological categories of the respective language. Phonological categories are language-specific, develop over years, and are updated throughout the individual's entire lifespan (Pierrehumbert, 2003). In speech perception, frequencies determine the emergence of phonological categories. They also form the basis for speech production, because a given category will be produced only when a critical number of perceived exemplars is stored in the perceptual space.

3.2 Prosodic targets in perceptual space

The model proposed by Guenther (Guenther, 1995; Guenther et al., 1998) posits that speech production is constrained by auditory and perceptual requirements. The only invariant targets of the speech production process, according to this model, are auditory perceptual targets. These invariant targets are characterized as multidimensional regions in the perceptual space. The relevance for speech production that the model attributes to perceptual constraints is the main distinguishing property that sets it apart from the mainly articulation-oriented models of speech production, such as the linguistic gestural model, as advocated by articulatory phonology (Browman and Goldstein, 1986, 1992), or the task dynamic model (Saltzman and Munhall, 1989).

Guenther's speech production model serves as the basis for a theory of speech motor control (Perkell et al., 1999, 2000). In this theory, the auditory-acoustic targets are interpreted as the basic programming units in speech motor control. Speech movements are planned with economy of effort, i.e. achieving sufficient perceptual contrast with minimal effort (Perkell et al., 2002). It is further hypothesized that motor-equivalent behavior may be the speech production counterpart of discrimination enhancement or perceptual sharpening in categorical perception, and of increased discrimination abilities with increased distance from perceptual prototypes in the perceptual magnet effect (Kuhl, 1991; Lacerda, 1995; Guenther and Gjaja, 1996).

Based on this model, we have proposed that speech movements in the prosodic domain are interpreted as prosodically relevant gestures that are planned to reach and traverse perceptual target regions (Dogil and Möbius, 2001; Möbius and Dogil, 2002). The targets are characterized as multidimensional regions in the perceptual space. Gestures that are successfully executed by the speaker produce acoustic realizations of perceptually relevant prosodic categories, such as those predicted by intonational phonology. Examples of mapping relations between the reference frame, i.e. the target regions, and prosodic gestures were also discussed (Dogil and Möbius, 2001).

3.3 Exemplar-theoretic interpretation of prosodic targets

It has been claimed that internal phonemic models emerge from storing in memory representations of large numbers of perceived acoustic realizations (Johnson, 1997; Pierrehumbert, 2001a). There is evidence that what is used in speech perception is these exemplars themselves, including their phonetic detail, rather than more abstract representations built from the exemplars. We have proposed that accumulations of perceived exemplars implicitly define prosodic target regions that are used in speech production (Schweitzer and Möbius, 2003). Under this assumption, prosodic categories are characterized by regions with an increased density of exemplars in perceptual space.

In speech *perception*, new tokens are perceived in identification tasks as belonging to the category that comprises the highest number of similar exemplars. Discrimination sensitivity depends on the local variation in the number of exemplars from competing categories (Lacerda, 1995). This model can account for the warping of the perceptual space attributed to the perceptual magnet effect as well as for categorical perception effects: peaks in discrimination sensitivity are expected between overlapping categories, whereas lower discrimination sensitivity is expected inside a category.

In speech *production*, exemplars can serve as perceptual targets. This account is compatible with the concept of perceptual target regions if one assumes that the accumulation of exemplars implicitly defines a corresponding region in perceptual space. Thus, the speaker has access to stored representations of prosodic events, including their tonal and temporal structure, that serve as a reference in speech production. The z-score measure introduced below (Section 3.3.2) requires that the listener also has access to a large number of acoustic realizations of segments occurring in different prosodic contexts.

In Section 3.3.1, we briefly review the key assumptions made by proponents of Exemplar Theory, inasmuch as they are relevant for an exemplar-theoretic interpretation of prosodic targets. Experimental evidence for the viability of this interpretation is presented in Section 3.3.2.

3.3.1 Exemplar Theory

Exemplar Theory assumes that speech perception and production are closely linked to each other in a perception-production loop. All percepts of speech events are stored in memory as exemplars in a perceptual space. This space can be represented as a cognitive map comprising many dimensions, which encode the phonetic and phonological properties of the exemplars. Percepts of nearly identical instances are located on the map in close vicinity to each other, whereas percepts of less similar instances are located in different regions. Thus, perceived realizations of speech events form clouds of exemplars on the map (Pierrehumbert, 2001b). These exemplar clouds represent the categories of a given language. Within each category the distribution of exemplars indicates the range of variation among the parameters which characterize the respective category. The optimal location of an exemplar prototype does not have to be represented by an existing exemplar token (Pierrehumbert, 2001a). Optimal locations may represent idealized, abstract prototypes.

A category label is assigned to each coherent exemplar cloud. The cognitive system assumed by Exemplar Theory can thus be described as a mapping from locations in the perceptual space onto labels of the language's category system.

The labels form their own level of representation; they may be regarded as functional links to other levels. Because each perceived exemplar is stored in memory, categories will emerge which may comprise many or just a few tokens. However, each exemplar will remain in memory only for a limited period of time (Goldinger, 1997). Exemplars that are not refreshed or activated by means of processing new perceptions by the internal analysis-by-synthesis (Figure 1), will be removed from memory (memory decay). In this way, entire categories may disappear from perceptual space. New categories may emerge from the perception of a sufficient number of stimuli for which no appropriate category exists yet in the perceptual space. It is assumed that a critical number of such exemplars must be perceived to form a new category (Pierrehumbert, 2003). Frequency of occurrence, or frequency of experience, is thus a crucial factor in the perception and production of speech. Given the perception-production link, the listener's ability to produce new categories is a function of the number and quality of input exemplars.

3.3.2 Experimental evidence

Prosodic speech events are perceived by the listener and stored as exemplars in the perceptual space according to the category labels assigned to them. Internal models of prosodic categories emerge from these stored exemplars; a sufficiently large number of perceived acoustic realizations is required for the emergence of new categories. The experimental results presented below support these hypotheses.

In a set of experiments aimed at exploring the compatibility of Guenther and Perkell's speech production model with an exemplar-theoretical view, the target regions of frequent vs. infrequent syllables were investigated (Schweitzer and Möbius, 2003, 2004). The introduction of frequency of occurrence as an independent factor was motivated by the idea that an exemplar-theoretical approach is compatible with the existence of a mental syllabary in which realizations of the most frequent syllables are stored (Levelt and Wheeldon, 1994). Syllables assumed to be stored in the syllabary have been shown to exhibit more coarticulation than rare syllables (Whiteside and Varley, 1998), which are assembled on-line from smaller units.

We will argue here that this finding would be predicted by an exemplar-theoretic interpretation of Guenther and Perkell's speech production model: infrequent units are represented by considerably fewer exemplars; for the most infrequent units, there may be no exemplars stored at all. This implies that no target regions are established for infrequent units, and that the speaker has to resort to smaller and more frequent units. In the following we present data from an experiment on syllable durations that reveal differences in the production of very frequent and very infrequent syllables (Schweitzer and Möbius, 2004).

Based on the finding of different degrees of coarticulation as a function of frequency of occurrence (Whiteside and Varley, 1998), we expected to find a similar frequency effect in the temporal domain as well. We demonstrated experimentally that the realization of a unit (i.e. segment or syllable) relative to the temporal target region can be modeled by using unit-specific mean duration and standard deviation (Schweitzer and Möbius, 2004).

We started from the hypothesis that there is less variation for infrequent syllables than for frequent syllables when syllable duration z-scores are compared to the mean segment duration z-scores of the involved segments.[6] This hypothesis was motivated by the assumption that, as the speaker has no access to syllable-specific mean duration and standard deviation for infrequent syllables, lengthening or shortening is planned for each segment relative to the pertinent segment mean and standard deviation. For frequent syllables, on the other hand, a sufficient number of representations is available to plan lengthening or shortening relative to the mean and standard deviation of the stored representations of the syllables in question.

We examined syllable realizations in a large (160 min., 95 000 phone realizations) single-speaker speech corpus (Schweitzer et al., 2003). This corpus had been designed to maximize coverage of phoneme-phoneme combinations, and therefore exhibits an unusual syllable frequency distribution with a disproportionately large number of instances of certain otherwise infrequent syllables. The frequency classification of the syllables was based on probabilistic syllable classes induced from multivariate clustering (Müller et al., 2000), which permits the estimation of the theoretical probability even for unseen syllables. From all syllable types which occurred more than 20 times in our database, we chose the 16 most infrequent types and the 114 most frequent types, corresponding to estimated probabilities of less than 0.00005 and more than 0.001, respectively. These added up to 12 278 realizations of frequent types and 471 realizations of infrequent types.

We calculated linear regression models for both sets separately, using the syllable z-score as the predictor variable and the mean segment z-score for the syllable as the predicted variable. The residual standard errors were 0.400 and 0.365 for frequent and infrequent syllables respectively, indicating stronger variation for the frequent syllables (Figure 8). The Bartlett test confirmed that the difference in variation for the residuals was significant at $p<0.0001$.

The results thus support our hypothesis that the temporal planning of a syllable is more directly predictable from the segmental durations if the syllable is of a frequent type than if the syllable is of a rare type. The relationship between syllable z-scores and the z-scores of the corresponding segments is significantly stronger for infrequent than for frequent syllables. We claim that this is due to the fact that infrequent syllables have to be assembled from smaller units because they are not represented by a sufficient number of exemplars to establish a syllable-level target region.

[6] The z-score indicates by how much a particular segment deviates from the phoneme's mean duration. Formally, the z-score of a segment p_i is the factor that has to be applied to the corresponding phoneme's standard deviation $\sigma(p)$ such that together with the phoneme's mean $\mu(p)$ it sums up to the observed segment duration:
$duration(p) = \mu(p) + z\text{-}score(p_i) * \sigma(p)$.

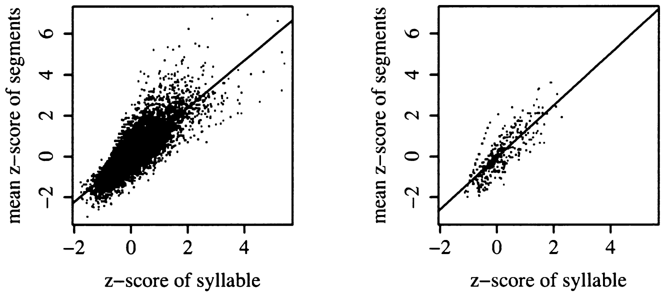

Figure 8: Mean duration z-scores of segments within a syllable plotted against duration z-score of the syllable for frequent (left panel) and infrequent (right panel) syllables

4 Conclusion

We have argued that phonetic categories in the segmental and in the prosodic domain emerge from probability distributions corresponding to regions in the parametric phonetic space. Our proposal integrates the concept of perceptual target regions (Guenther et al., 1998; Perkell et al., 2000; Dogil and Möbius, 2001) with predictions made by Exemplar Theory (Johnson, 1997; Pierrehumbert, 2001a). We suggest that categories emerge as a consequence of applying an internal analysis-by-synthesis process, which sends the underspecified category through a loop in which the category is internally synthesized by the hearer. The result of the synthesis is then internally compared to the original signal in its original context. The comparison is performed in the phonetic space that is neutral to the hearer's and speaker's articulatory-acoustic space. Internal prosodic models are therefore assumed to be defined in the perceptual reference space.

Frequency of occurrence, or frequency of experience, is an important factor for the establishment of a prosodic category. Every time the realization of a category is perceived, it is processed by the analysis-by-synthesis loop. This entails that the cues which have a high frequency of occurrence in a specific context will be represented by a large number of exemplars in the hearer's perceptual reference space. The plausibility of these assumptions was tested in an experiment whose results showed the effect of frequencies of occurrence on the temporal planning of syllables. We have further argued that accumulations of exemplars develop categorical properties and exert magnet effects.

Current probabilistic accounts emphasize the explanatory importance of frequency, and Exemplar Theory is perhaps the best instantiation of a probabilistic model that is linguistically explanatory. It places the number and quality of percepts into the core of the model. Given the perception-production link, the listener's ability to produce new categories is therefore a function of the frequency and properties of input exemplars.

Phonetic research provides elaborate methods of investigating the categorical quality of exemplars of speech events. In this paper we demonstrated the application of two experimental paradigms for testing the categorical status of elements of the prosodic structure: the categorical perception paradigm and the perceptual magnet effect paradigm. Both classical paradigms appear to be useful diagnostic testing tools in the domain of intonation.

In our presentation of the CP and PME experiments, we have focused on methodological issues, such as experimental design; stimulus generation; series of subtests including identification, goodness rating and discrimination; and the evaluation and interpretation of the experimental results. We have stressed the importance of statistical analysis methods and the relevance of Signal Detection Theory for a proper interpretation of the results.

Future work should address the refinement of details of the experimental design. For instance, varying the number and names of the answer categories offered to the participants of the perception tasks has recently been shown to affect the results: in their experiments on the perception of boundary tones in Dutch, van Heuven and Kirsner (2004) asked the listeners to respond by choosing from "command vs. no command" and "question vs. no question" in a two-category design, and from "command vs. condition vs. question" in a three-category design, aiming at a better perceptual definition of the category boundary or boundaries.

References

Batliner, A. (1989): Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorien. In: H. Altmann, A. Batliner, and W. Oppenrieder (eds.): Zur Intonation von Modus und Fokus im Deutschen. Tübingen: Niemeyer, 111–162.

Browman, C. and L. Goldstein (1986): Towards an articulatory phonology. Phonology Yearbook 3, 219–252.

Browman, C. and L. Goldstein (1992): Articulatory phonology: An overview. Phonetica 49, 155–180.

Bruce, G. (1977): Swedish Word Accents in Sentence Perspective. Lund: Gleerup. Travaux de l'Institut de Phonétique XII.

Büring, D. (1997): The Meaning of Topic and Focus – The 59th Street Bridge Accent. New York: Routledge.

Cinque, G. (1993): A null theory of phrase and compound stress. Linguistic Inquiry 24(2), 239–297.

Clements, G. N. and K. Ford (1979): Kikuyu tone shift and its synchronic consequences. Linguistic Inquiry 10, 179–210.

Dogil, G. (2003): Understanding prosody. In: G. Rickheit, T. Herrmann, and W. Deutsch (eds.): Psycholinguistik – Ein internationales Handbuch / Psycholinguistics – An International Handbook. Berlin: de Gruyter, 544–565.

Dogil, G., J. Kuhn, J. Mayer, G. Möhler, and S. Rapp (1997): Prosody and discourse structure. In: A. Botinis, G. Kouroupetroglou, and G. Carayiannis (eds.): Intonation: Theory, Models and Applications – Proceedings of an ESCA Workshop, Athens, Greece, 99–102.

Dogil, G. and B. Möbius (2001): Towards a model of target oriented production of prosody. In: Proceedings of the European Conference on Speech Communication and Technology, Aalborg, Denmark, vol. 1, 665–668.

Féry, C. (1993): The Meaning of German Intonational Patterns. Tübingen: Niemeyer.

Goldinger, S. D. (1997): Words and voices – Perception and production in an episodic lexicon. In: K. Johnson and J. W. Mullennix (eds.): Talker Variability in Speech Processing. San Diego: Academic Press, 33–66.

Goldsmith, J. A. (1990): Autosegmental and Metric Phonology. Oxford: Blackwell.

Guenther, F. H. (1995): A modeling framework for speech motor development and kinematic articulator control. In: Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm, vol. 2, 92–99.

Guenther, F. H. (2003): Neural control of speech movements. In: N. O. Schiller and A. S. Meyer (eds.): Phonetics and Phonology in Language Comprehension and Production. Berlin: Mouton de Gruyter, 209–239.

Guenther, F. H. and M. N. Gjaja (1996): The Perceptual Magnet Effect as an emergent property of neural map formation. Journal of the Acoustical Society of America 100, 1111–1121.

Guenther, F. H., M. Hampson, and D. Johnson (1998): A theoretical investigation of reference frames for planning of speech movements. Psychological Review 105, 611–633.

Heeger, D. (2003): Signal Detection Theory. www.cns.nyu.edu/˜david/sdt/ sdt.html, last accessed 03/2005.

Helmholtz, H. L. F. v. (1863): Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik. Braunschweig: Vieweg.

Hermes, D. J. and J. C. van Gestel (1991): The frequency scale of speech intonation. Journal of the Acoustical Society of America 90, 97–102.

House, D. (1996): Differential perception of tonal contours through the syllable. In: Proceedings of the International Conference on Spoken Language Processing, Philadelphia, vol. 1, 2048–2051.

Iverson, P. and P. K. Kuhl (1995): Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. Journal of the Acoustical Society of America 97, 553–562.

Johnson, K. (1997): Speech perception without speaker normalization: An exemplar model. In: K. Johnson and J. W. Mullennix (eds.): Talker Variability in Speech Processing. San Diego: Academic Press, 145–165.

Kohler, K. J. (1987): Categorical pitch perception. In: Proceeding of the 11th International Congress of Phonetic Sciences, Tallinn, 331–333.

Kohler, K. J. (1990): Macro and micro F0 in the synthesis of intonation. In: J. Kingston and M. E. Beckman (eds.): Papers in Laboratory Phonology. Cambridge: Cambridge University Press, vol. 1, 115–138.

Kuhl, P. K. (1991): Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. Perception and Psychophysics 50, 93–107.

Lacerda, F. (1995): The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In: Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm, vol. 2, 140–147.

Ladd, D. R. and R. Morton (1997): The perception of intonational emphasis: Continuous or categorical? Journal of Phonetics 25, 313–342.

Levelt, W. J. M. and L. Wheeldon (1994): Do speakers have access to a mental syllabary? Cognition 50, 239–269.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith (1957): The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54(5), 358–368.

Massaro, D. W. (1998): Categorical perception: Important phenomenon or lasting myth? In: Proceedings of the International Conference on Spoken Language Processing, Sydney, vol. 6, 2275–2278.

Möbius, B. and G. Dogil (2002): Phonemic and postural effects on the production of prosody. In: B. Bel and I. Marlien (eds.): Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence, 523–526.

Moulines, E. and F. Charpentier (1990): Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9, 453–467.

Müller, K., B. Möbius, and D. Prescher (2000): Inducing probabilistic syllable classes using multivariate clustering. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 225–232.

Perkell, J. S., F. H. Guenther, H. Lane, M. L. Matthies, P. Perrier, J. Vick, R. Wilhelms-Tricarico, and M. Zandipour (2000): A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. Journal of Phonetics 28(3), 233–272.

Perkell, J. S., F. H. Guenther, H. Lane, M. L. Matthies, J. Vick, and M. Zandipour (2001): Planning and auditory feedback in speech production. In: B. Maassen, W. Hulstijn, R. D. Kent, H. F. M. Peters, and P. H. H. M. van Lieshout (eds.): Proceedings of the 4th International Speech Motor Conference (Nijmegen), 5–11.

Perkell, J. S., M. Zandipour, M. L. Matthies, and H. Lane (1999): Articulatory kinematics: Preliminary data on the effects of speaking condition, articulator and movement type. In: Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, vol. 3, 1773–1776.

Perkell, J. S., M. Zandipour, M. L. Matthies, and H. Lane (2002): Economy of effort in different speaking conditions I: A preliminary study of intersubject differences and modeling issues. Journal of the Acoustical Society of America 112(4), 1627–1641.

Pierrehumbert, J. (2000): The phonetic grounding of phonology. Les Cahiers de l'ICP, Bulletin de la Communication Parlée 5, 7–23.

Pierrehumbert, J. (2001a): Exemplar dynamics: Word frequency, lenition and contrast. In: J. Bybee and P. Hopper (eds.): Frequency and the Emergence of Linguistic Structure, Amsterdam: Benjamins, 137–157.

Pierrehumbert, J. (2001b): Stochastic phonology. GLOT 5(6), 1–13.

Pierrehumbert, J. (2003): Probabilistic phonology: Discrimation and robustness. In: R. Bod, J. Hay, and S. Jannedy (eds.): Probability Theory in Linguistics. Cambridge, MA: MIT Press, 177–228.

Remijsen, B. and V. van Heuven (1999): Gradient and categorical pitch dimensions in Dutch: Diagnostic test. In: Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, vol. 3, 1865–1868.

Repp, B. (1984): Categorical perception: Issues, methods, and findings. In: N. J. Lass (ed.): Speech and Language: Advances in Basic Research and Practice. New York: Academic Press, vol. 10, 243–335.

Saltzman, E. L. and K. G. Munhall (1989): A dynamical approach to gestural patterning in speech production. Ecological Psychology 1(4), 333–382.

Schiefer, L. and A. Batliner (1991): Order effect and the order of accents. In: Proceedings of the 12th International Congress of Phonetic Sciences, Aix-en-Provence, vol. 3, 86–89.

Schneider, K. and B. Lintfert (2003): Categorical perception of boundary tones in German. In: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 631–634.

Schweitzer, A., N. Braunschweiler, T. Klankert, B. Möbius, and B. Säuberlich (2003): Restricted unlimited domain synthesis. In: Proceedings of Eurospeech-2003, Geneva, 1321–1324.

Schweitzer, A. and B. Möbius (2003): On the structure of internal prosodic models. In: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 1301–1304.

Schweitzer, A. and B. Möbius (2004): Exemplar-based production of prosody: Evidence from segment and syllable durations. In: Speech Prosody 2004, Nara, Japan, 459–462.

Stevens, K. N. (1989): On the quantal nature of speech. Journal of Phonetics 17, 3–45.

van Heuven, V. J. and R. S. Kirsner (2004): Phonetic or phonological contrasts in Dutch boundary tones? Linguistics in the Netherlands 21, 102–113.

Whiteside, S. P. and R. A. Varley (1998): Dual-route phonetic encoding: Some acoustic evidence. In: Proceedings of the 5th International Conference on Spoken Language Processing, Sydney, vol. 7, 3155–3158.

Wickens, T. D. (2002): Elementary Signal Detection Theory. Oxford University Press.